Three Criteria for Selecting Variables in the Construction of Near-Optimal Decision Trees

AUTHOR(S):

MIYAKAWA, Masahiro; OTSU, Nobuyuki

数理解析研究所講究録
第 731 巻 1990 年 238-249

Three Criteria for Selecting Variables in the Construction of Near-Optimal Decision Trees

Masahiro MIYAKAWA and Nobuyuki OTSU
(宮 川 正 弘)　　(大 津 展 之)

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Japan 305

Abstract

For converting a decision table to a near-optimal decision tree in the sense of the minimal number of nodes of the tree, we propose three criteria for variable selection: Γ_A, Γ_H and Γ_D from three different standpoints of combinatorial, entropy and discriminant analyses. First we examine "static" behaviors of the criteria, e.g. rejection of nonessential variables (*nev*-free), selection of a totally essential variable (*tev*-bound) and selection of a quasi-decisive variable (*qdv*-bound). It is shown that Γ_A is *nev*-free, *tev*-bound but not *qdv*-bound, while Γ_H and Γ_D have the complemental properties. An experiment to evaluate the performance of the criteria shows that each of the criteria gives good near-optimum trees, indicating that Γ_D and Γ_H are practically comparative and Γ_A is slightly better. All the criteria require at most $O(L^2 2^L)$ operations with $O(L2^L)$ storage, where L is the number of variables of the input table.

1. Introduction

A decision table is a list of "rules". A rule consists of a pair of "logical conditions" and an "action" which specifies the action to be executed when the conditions are satisfied. Taking Boolean values for both conditions and actions, it represents a Boolean function; taking feature vectors for conditions and object names for actions, it gives a description of a pattern. The occurrence of a decision table is rather wide. Identification of a specimen in biological science is a practical example of decision table. If we count its use as a conceptual tool, then it appearance is even wider. For example, a typical situation in knowledge engineering or even a program, e.g. so called Janov scheme, can also be schematized in decision table terms. An important feature of a decision table is that it specifies only a very restricted part of logical possibilities and leaves the remaining part as "don't cares" or unconcerned (in other words most decision tables specify partial functions).

In a typical use of a decision table to identify an unknown "input object", one tests each single property in sequence until the object is determined uniquely. This procedure is called a *sequential test procedure* and is conveniently represented by a decision tree [ReS67]. In most practical cases a decision may be made by testing only some of the properties. In addition there are fairly large tables (e.g. having dozens number of properties) which are unable to be manually converted into efficient trees. Thus an automatic conversion of a table into an optimum or near-optimum tree is most desirable [MTG81] and the central problem being to determine the order of testing of the properties.

It is known that the construction problems of various optimum decision trees are NP-complete [HyR76, Mor82] when a given input table is not in "expanded" form (most frequent in practice as we mentioned before). Several heuristics of constructing near-optimum trees have been proposed; mostly for treating the don't care entries (cf. [Ver72]). Given an L-variable table, one can construct an optimum, i.e. a minimum-cost tree by applying a dynamic programming technique which *always* requires $O(L3^L)$ operations (comparisons) with $O(3^L)$ storage [Bay73, ScS76]. On the other hand, a simple top-down heuristic method employing a successive variable selection, viz. VSM (variable selection method) can construct a near-optimum tree in far less operations if an efficient selection is guaranteed. Moreover, it has an advantage of applicability to a most practical case of partial functions.

In this paper we adopt the number of internal nodes of a tree as a cost of a tree (i.e. as an optimality criterion) and propose three criteria for the VSM: Γ_A from the combinatorial standpoint, Γ_H from the entropy standpoint and Γ_D from the discriminant analysis standpoint. Naturally we want to determine their performance comparing with optimal one and also the

best heuristic among the three. Can we prove this by some means? As a first approach we investigate their behaviors in typical situations, namely we check rejection of nonessential variables (*nev*-free) and selection of a totally essential variable or a quasi-decisive variable (*tev*- or *qdv*-bound). Note that *nev* is a worst variable while *tev* and *qdv* are optimal ones. We show that Γ_A is *nev*-free and *tev*-bound but not *qdv*-bound, while Γ_H and Γ_D have just the complement properties. Thus, as a next step, this leads us to conduct an experiment research to compare the performance. It shows that the criteria actually give near optimum trees. Also the performance of Γ_D and Γ_H practically coincides (1.05 in the average compared with optimum one) and Γ_A is slightly better (1.03 by the previous measure). All the criteria require at most $O(L^2 2^L)$ operations (bit-comparisons or integer-additions) with $O(L2^L)$ storage (this is a polynomial bound if we take 2^L size of an expanded table into account).

The construction of a tree having a minimum number of nodes has been received less attention to compared with that of a tree having a minimum average path lengths [Gar72, MiS80, MTG81, Miy85]. However, the minimum cost is an invariance of the table, representing the minimum number of "dividing" of the table necessary to decompose it into constant tables; a quantity inherently related to a complexity of the table [Bud85, Lov85, Weg84]. This problem is also directly related to the design of PLA (programmable logic array) in which it is important to obtain complement of a logical function in the form of as few product terms as possible [Cha87, Sas85].

2. Definitions and Preliminaries

Let us denote L *properties* simply by numbers $\{1,\ldots,L\}$. Let $\{a_1,\ldots,a_K\}$ be the set of K *actions*. Assume that each property i takes, for simplicity, the binary value $x_i = 0$ or 1. We are given a mapping $f : \{0,1\}^L \to \{a_1,\ldots,a_K\}$, called an L-ary-K-action *decision table*, or simply a *table*, which maps the values of the properties into the actions. Let $\boldsymbol{x} = (x_1 \cdots x_L) \in \{0,1\}^L$. A pair $(\boldsymbol{x}, f(\boldsymbol{x}))$ is called a *rule*. Often we denote a rule simply by a vector $\boldsymbol{x}$ since it uniquely determines a rule. We treat a table f as the set of all 2^L rules. In Table 1 we give an example of a table in reduced form (a) and in expanded form (b).

2.1. Subtables and fixations

Given an *initial* table f, a *subtable* is a restriction

$$f(x_1 \ \cdots \ x_{i_1-1} s_1 x_{i_1+1} \cdots x_{i_h-1} \cdots s_h x_{i_h+1} \cdots x_L) = f|(x_{i_m} = s_m \text{ for } m = 1,\ldots,h)$$

Table 1: A decision table: (a) reduced form, (b) expanded form.

1	2	3	action
–	1	0	a
1	0	–	a
0	0	0	b
1	1	1	b
0	–	1	c

123	action
000	b
001	c
010	a
011	c
100	a
101	a
110	a
111	b

which is an $(L-h)$-ary function. The variables x_{i_m}, $m = 1,\ldots,h$, $0 \le h \le L$ are called *fixed* (to s_m; $s_m = 0$ or 1), and the remaining variables are *free*. A subtable consists of all vectors some (say h) of their elements equal fixed constants. The number $L-h$ of free variables is the *arity* of the subtable. Sometimes it is convenient to show it explicitly like $L-h$-subtable (thus the initial table is an L-subtable while a rule is a 0-subtable). Since the initial table consists of 2^L rules, we have 2^{2^L} different subsets, among them 3^L subsets are subtables.

The sole type of restrictions $f|(x_{i_m} = s_m \text{ for } 1 \le m \le h)$ we deal with is called a *fixation of* f and denoted by $fi_1^{s_1}\cdots i_h^{s_h}$ or simply $f\alpha$, where α denotes a catenation $i_1^{s_1}\cdots i_h^{s_h}$. The null fixation λ corresponds to the initial table f. This notation conveniently represents the successive generation of subtables in the tree by means of "dividing" of a table, which is described below.

2.2. Decision trees and Variable Selection Method (VSM)

A *decision tree* [Miy85] for f is a binary tree associated with each internal node two objects: a subtable and a variable (called a *test variable*), and with each leaf a *decision action* according to the following algorithm:

If f is a constant (i.e. f consists of a single action) then the decision tree for f is simply a single leaf with the decision action. Otherwise it is a tree with a root associated with the subtable f and with any its free variable i as the test variable, having the left and right subtrees corresponding to the subtables fi^0 and fi^1, respectively (the edges leading to them are labeled by 0 and 1, respectively).

Thus every tree we deal with is an *extended binary tree* [9, p.399], each node of it having exactly one *in-*

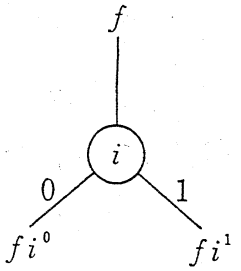

Figure 1: Dividing a table f by a variable i.

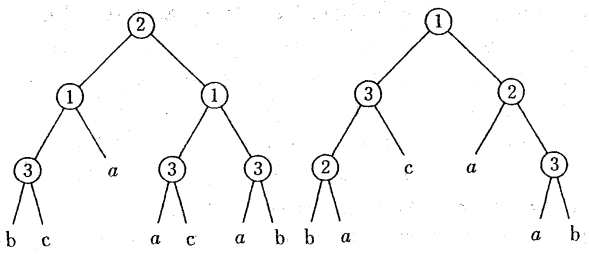

Figure 2: Example trees: (a) non-optimal. (b) optimal.

edge (except the root R which has no in-edge) and either zero or two *out-edges*. *External nodes* (leaves) are those with no out-edges. The remaining nodes are called *internal*. We call a decision tree simply a *tree*.

One may think that each rule x of a table f is identified as belonging to fi^0 or fi^1 according to $x_i = 0$ or 1, respectively, at each internal node where the test variable i is located. A path of the tree represents successive such fixations. Thus a tree can be considered as a device to determine values of a given function by means of successive fixations. We denote a tree by T, or $T[f]$ when it is necessary to indicate the initial table.

Given a criterion to choose a test variable, one can construct a tree by consistently selecting test variables according to the criterion. This general algorithm for constructing a tree is called a *Variable Selection Method* (VSM). The basic operation of making $L-1$-subtables fi^0 and fi^1 from f is called "dividing" of a table (this corresponds to refining the action set in the table).

A VSM ends in $O(2^L)$ operations, since possible number of such dividing of a table is at most $2^L - 1$.

2.3. Minimum decision trees

For an L-ary table f, there can be at most $\prod_{i=1}^{L} i^{2^{i-1}}$ equivalent trees [ScS76]. To construct "efficient" trees we adopt the number of internal nodes of a tree T as a *cost* of a tree and denote it by $|T|$. Since no test is required at leaves, $|T|$ is the cost for representing the whole test procedures of T, assuming that the determination of the value of each property incurs a uniform cost. We call a tree *minimum* (optimum) when its cost is a minimum among all trees corresponding to the table f. In Fig.2 we indicate a tree and a minimum tree for the table of Table 1.

2.4. Nonessential, totally essential and quasi-decisive variables

Before going to variable selection criteria, we introduce some notions concerning "good" variables. To do this we need a notation to represent fixed and free variables of a subtable. Our discussion usually concerns an arbitrarily fixed subtable f and its direct subtables fi^s for $s = 0, 1$. Let f have all h variables denoted by $1, \ldots, h-1$ and j $(i \neq j, i = 1, \ldots, h-1)$. Let $u = u(1) \ldots u(h-1)$, $u(i) = i$, $i = 1, \ldots, h-1$ denote a sequence (i.e. catenation) of variables $1, \ldots, h-1$. Let $x = x(1) \ldots x(h-1)$, $x(i) = 0$ or 1 for $i = 1, \ldots, h-1$, denote a bit sequence of length $h-1$. Let us abbreviate a fixation $u(1)^{x(1)} \ldots u(h-1)^{x(h-1)}$ by u^x for simplicity. Now, let $u^x j$ represent a vector in which the variables in u are fixed to the values x and a variable j is free. Finally, let us denote a 1-subtable called a *j-pair* by $f(u^x j) \equiv f u^x (u^x j)$, which consists of a pair of rules of f: $(u^x j^0, f(u^x j^0))$ and $(u^x j^1, f(u^x j^1))$. If two actions of a j-pair coincide, it is a *constant j-pair*, otherwise it is a *nonconstant j-pair*. Each fixation u^x which gives constant or nonconstant j-pair is called *inactive* or *active* fixation for j, respectively (hereafter only x is indicated instead of u^x since u is determined as the "complement" sequence of j).

Example 2.1. For $x = 10$ and $u = 23$ the fixation u^x denotes $2^1 3^0$. Then 1-pair $f(2^1 3^0 1) = f 2^1 3^0 (2^1 3^0 1)$ in Table 2.1 is a constant 1-pair, while $f(2^0 3^0 1) = f 2^0 3^0 (2^0 3^0 1)$ is a nonconstant 1-pair. Hence the fixation $2^0 3^0$ is active while $2^1 3^0$ is inactive.

A variable i is *nonessential* in f if each fixation x for i is inactive, i.e. $f(u^x i) = a_j$ for each x (a constant action a_j depends on x). Also, a variable i is *totally essential* if each fixation x for i is active, i.e.

$f(u^x i) \neq const.$ for each x. When all variables are nonessential then f is a constant. Nonessential variable and totally essential variable are abbreviated to *nev* and *tev*, respectively. A minimum tree doesn't have *nevs* as test variables [15]. A *tev* i is an optimum variable [6], that is, the tree becomes minimum if we choose i as a test variable and make its left and right subtrees minimum for fi^0 and fi^1, respectively. Note that there could be a case that some of the rules are never executed actually (e.g. they could be implied by others in practical tables; this can be handled by assigning "probability" 0 to them). Then even "essential variables" need not be tested at all in the tree. So in this case the notion of "essentiality" should be modified in an appropriate way. Throughout this paper we assume that each rule is executable, i.e. the probability of each rule is not 0.

There is another optimal variable. A variable i is *decisive* if $fi^0 = a$ (*cost.*), $fi^1 = b$ (*const.*) and $a \neq b$. A variable i is *quasi-decisive* (abbreviated to *qdv*), if $fi^s = const.$ and $fi^{\bar{s}} \neq const.$ for strictly either one of $s = 0$ or 1 [14]. A *qdv* is an optimal variable with respect the cost defined in this paper (however, it is not an optimal variable in general with respect to another one (e-cost)) [14].

Therefore, it is desirable for a criterion to reject *nevs* and select a *tev* or a *qdv* whenever they exist. Let us call a criterion *nev-free* if it selects no *nevs*. Again, let us call a criterion *tev-bound* or *qdv-bound* if it selects a *tev* or a *qdv* whenever they exists.

In the next section we propose three VSM criteria and examine above properties for them. This is important for understanding the performance of the criteria, since it is extremely difficult to show something formally about the performance of this kind of heuristics.

3. Three criteria for variable selection

We shall introduce three criteria for variable selection to construct an efficient tree in the sense of minimal number of internal nodes (i.e. dividing). Our basic strategy is to "produce constant tables as fast as possible", since no more dividing is necessary for constant tables. To do this we introduce three measures of "distance" between a table and a constant from three different standpoints. Then our criteria simply select such i that the "distance" between fi^s and a constant function for both $s = 0$ and 1 become a minimum among all variables.

The first criterion Γ_A is implicit in some literature [23,19], approaching the problem from combinatorial analysis. The second criterion Γ_H is presented in [7,13,17]. Also there are several literature in which the notion of entropy is applied for the problem in other context (mostly with the treatment of "don't care" symbols "-") [6]. The entropy approach "views" the occurrences of the actions as stochastic events. The third criterion Γ_D is new [13] and from the discriminant analysis standpoint [4], which is related to multivariate data analysis. In this framework, we use the notion of "mean value" of the variable x_i with respect to the action a_j (denoted by μ_j^i) as well as the total mean value (denoted by μ_T^i whose value is always 1/2 since we have the same number of rules having $x_i = 0$ and $x_i = 1$), treating the values 0 and 1 as real numbers.

Before giving a detailed explanation of the criteria, we give notations and equations. We denote by N_j the number of the occurrences of the action a_j in f, by $N_j^{i^s}$ the number of rules $(\boldsymbol{x}, f(\boldsymbol{x}))$ such that $f(\boldsymbol{x}) = a_j$ and $x_i = s$ for $s = 0, 1$. Further, let $class(j)$ denote the set of vectors corresponding to action a_j. There hold the following equations.

1) $N_j^{i^0} + N_j^{i^1} = N_j$,
2) $\sum_j N_j^{i^s} = N^{i^s} = 2^{L-1} = N/2,\ s = 0, 1$,
3) $\sum_j N_j = N = 2^L$,
4) $p_j^{i^s} := N_j^{i^s}/N^{i^s} = 2N_j^{i^s}/N,\ s = 0, 1$,
5) $w_j := N_j/N,\ \sum_j w_j = 1$,
6) $\mu_j^i := E_j\ x_i = \sum_{class(j)} x_i/N_j = N_j^{i^1}/N_j$,
7) $\mu_T^i := E\ x_i = \sum x_i/N = \sum_j N_j^{i^1}/N = 1/2$;
$\mu_T^i = \sum_j w_j \mu_j^i$.

Now we present the three criteria together with their ranges for all tables.

Activity criterion Γ_A. Define A_i to be the number of active i-pairs of f. An active i-pair can be considered as a logical unit to be separated by dividing a table (cf. Fig 3). Then $A := \sum_i^n A_i$ represents the total number of i-pairs of f. Since A_i active i-pairs disappear by dividing f by i, we select a variable i which have a maximum number of A_i among all variables.

Lemma 3.1. *The number of active i-pair A_i ranges $0 \leq A_i \leq 2^{L-1}$. The best value $A_i = 2^{L-1}$ is attained if and only if i is a totally essential variable of f, and the worst value $A_i = 0$ if and only if i is a nonessential variable of f.*

Proof. Obvious from the definition. □

Entropy criterion Γ_H. The actions a_j in a table f can be considered to occur with the probabilities $w_j = N_j/N,\ j = 1, \ldots, K$. Therefore the nondeterminacy (ambiguity) among them can be measured by the entropy defined by

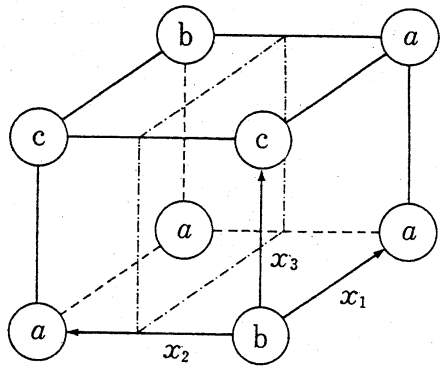

Figure 3: Dividing by a variable 2 reduces 2 active 2-pairs.

$$H(f) := -\sum_{j=1}^{K} w_j \log\ w_j. \tag{1}$$

The ambiguity remaining after testing a variable i may be defined as the average ambiguity of the two fixations fi^s for $s = 0$ and 1. The VSM procedure can be seen as a process of perpetual increase of the determinacy (equivalently, decrease of the ambiguity) of the tables in each fixation until finally we get all tables completely determined. Thus the amount of information obtained by testing the variable i may be defined by

$$H(f)-H(f|i) = H(f)-(-\sum_{s=0,1} w(fi^s)\sum_{j=1}^{K} p_j^{i^s} \log\ p_j^{i^s}), \tag{2}$$

where $w(fi^s) := p(fi^s)/p(f)$ denote the conditional probabilities $p(fi^s|f)$ of f^{i^s} under f, $p_j^{i^s}$ the conditional probabilities that an action a_j occurs under $x_i = s$, i.e. $p_j^{i^s} = p(a_j|fi^s)$ for $s = 0, 1$. Since the first term is a constant for all variables of f, only the second term is sufficient. Also, in our case the probability of an action is defined by the ratio of the number of rules having the action to the number of whole rules. Hence we have $w(fi^s) = 1/2$ for $s = 0, 1$. Thus, we select a variable i which has the least value of

$$H_i := -1/2 \sum_{s=0,1}\sum_{j=1}^{K} p_j^{i^s} \log\ p_j^{i^s}. \tag{3}$$

Lemma 3.2. *The entropy H_i ranges $0 \le H_i \le \log K$. The best value $H_i = 0$ is attained if and only if either f is a constant or i is a unique essential variable of f, and the worst value $H_i = \log K$ if and only if each action a_j occurs equiprobably for $j = 1, \ldots, K$ in both subtables fi^s for $s = 0$ and 1, i.e. $p_j^{i^s} = 1/K$ for $j = 1, \ldots, K$ and $s = 0, 1$ (equivalently, $N_j^{i^s} = N/(2K)$).*

Proof. Put $H_i = (-1/2)\sum_{s=0,1}\sum_{j=1}^{K} p_j^{i^s} \log\ p_j^{i^s} = 0$. Since the function $-p\log p$ $(0 \le p \le 1)$ is a non-negative convex function, $H_i = 0$ is attained in, and only in, the following four cases: $(p_j^{i^0} = 0$ or $p_j^{i^0} = 1)$ and $(p_j^{i^1} = 0$ or $p_j^{i^1} = 1)$ for each j. However, the case $p_j^{i^0} = 0$ and $p_j^{i^1} = 0$ for all j can be excluded, because this means that no action appear in f. Thus we have $p_j^{i^s} = 1$ for some j_s for $s = 0$ and 1, i.e. $fi^s = a_{j_s}$ (a constant) for $s = 0$ and 1. Thus, if $a_{j_0} = a_{j_1}$ then f is a constant, otherwise i is a unique essential variable of f. It is well-known that the unique maximum value of the entropy function (1) is attained iff the probabilities of the actions are uniform, i.e. $w_j = 1/K$. In equation (3) this should hold for both $s = 0$ and 1 (to assure this we assume that $2K$ divides N). □

Note 3.3. Under given occurrences of actions: $N_1, \ldots, N_K$ $(\sum_j N_j = N)$, the maximum value $H_i = -(1/N)\sum_j N_j \log N_j + \log N$ is attained if and only if i divides each N_j into equal halves for all j, i.e. $N_j^{i^0} = N_j^{i^1} = N_j/2$ for $1 \le j \le K$ (further when $N_j = N/K$ we have the worst value of the lemma).

Discriminant criterion Γ_D. In a decision table each variable x_i contributes to discriminate different actions. Hence one can measure "discriminating power" of a variable x_i from the standpoint of the discriminant analysis [4], which presents the following ratio of variances as its measure:

$$\eta_i^2 := \sigma_{Bi}^2/\sigma_{Wi}^2 \tag{4}$$

where σ_{Bi}^2 and σ_{Wi}^2 represent the between-action (interclass) and the within-action (intraclass) variances of the variable x_i, respectively, and defined by

$$\sigma_{Bi}^2 := \sum_j w_j(\mu_j^i - \mu_T^i)^2, \tag{5}$$

$$\sigma_{Wi}^2 := \sum_j w_j \cdot (1/N_j) \sum_{class(j)} (x_i - \mu_j^i).$$

To make the point of the theory clear, let us consider an example that we have a single decisive variable i which separates actions a and b, i.e. $fi^0 = a$ and $fi^1 = b$. The mean values of x_i with respect to the actions a and b is $\mu_a^i = 0$ and $\mu_b^i = 1$, respectively, discriminating the two actions precisely. If the two actions occur in both subtables, the values μ_a^i and μ_b^i are settled somewhere between 0 and 1, reflecting the mixed occurrences of the actions. Again, if the occurrences of the two actions are completely random ($x_i = 0$ and $x_i = 1$ occur the same number for both a and b), we have $\mu_a^i = \mu_b^i = 1/2$. This example illustrates that "mean values" are useful for measuring "discriminating ability" of a variable. One may wonder that the values 0 and 1 assumed by a variable are "nominal" entities only to be used to distinguish two different things. However, as is shown in the following Note 3.2 we can use 0 and 1 in place of any two different real numbers.

Note 3.4. The η^2 so defined by a ratio of two variances is invariant under an affine transformation of coordinate x to $y = b(x + a)$, i.e. shift a and scale factor b; in other words from $x = 0, 1$ to $y = ab$, $b(1 + a)$.

Since $\sigma_{Bi}^2 + \sigma_{Wi}^2 = \sigma_{Ti}^2$ and $\sigma_{Ti}^2 = 1/N \sum_{x \in I}(x_i - \overline{x})^2 = 1/4 = const.$ (since $\overline{x} = 1/2, N = 2^L$), the greater $\sigma_{B_i}^2$, the greater η_i, and the i which gives a maximum η_i also gives a maximum value for σ_{Bi}^2. Thus the interclass variance $\sigma_{B_i}^2$ alone represents the degree of separation of the classes. Hence we can take σ_{Bi}^2 as a selection criterion. Using $\sum w_j = 1$, finally we have:

$$\sigma_{Bi}^2 = \sum_j w_j(\mu_j^i)^2 - (\mu_T^i)^2 \quad (6)$$

$$= \sum_j w_j(\mu_j^i)^2 - 1/4 \quad (7)$$

$$= (1/N)\sum_j (N_j^{i^1})^2/N_j - 1/4 \quad (8)$$

(thus actually it suffices to calculate the first term).

Lemma 3.5. *The between-action variance σ_{Bi}^2 ranges $0 \le \sigma_{Bi}^2 \le 1/4$. The best value $\sigma_{Bi}^2 = 1/4$ is attained if and only if i divides the action set into two disjoint sets, i.e. the actions of fi^0 and fi^1 have no action in common, and the worst value $\sigma_{Bi}^2 = 0$ if and only if i divides each action class into two halves, i.e. $N_j^{i^0} = N_j^{i^1} = N_j/2$ (equivalently, $\mu_j^i = 1/2$) for each class j, $j = 1, \ldots, K$.*

Proof. Since $0 \le \mu_j^i \le 1$, we have $0 \le (\mu_j^i - 1/2)^2 \le 1/4$ and the maximum is attained if and only if $\mu_j^i = 0$ or 1.

Table 2: Values of the measures for the function.

variable	1	2	3
A_i	3	2	3
μ_a^i	3/4	2/4	1/4
μ_b^i	1/2	1/2	1/2
μ_c^i	0	1/2	1
σ_{Bi}^2	3/32	0	3/32
$p_a^{i^0}$	1/4	2/4	3/4
$p_b^{i^0}$	1/4	1/4	1/4
$p_c^{i^0}$	2/4	1/4	0
$p_a^{i^1}$	3/4	2/4	1/4
$p_b^{i^1}$	1/4	1/4	1/4
$p_c^{i^1}$	0	1/4	2/4
H_i	1.16	1.50	1.16

Since $\sigma_{Bi}^2 = \sum w_j(\mu_j^i - 1/2)^2 \le \sum w_j \cdot 1/4 = 1/4$ and this is attained if and only if $\mu_j^i = 0$ or 1 for all j. From $\sigma_{Bi}^2 = \sum_j w_j(\mu_j^i - \mu_T^i)^2 = 0$ we have $\mu_j^i = \mu_T^i = 1/2$ for all j. □

We note that the same situation that i divides each action class into two halves gives the worst values for both H_i and σ_{Bi}^2 (cf. Note 3.4). Also note that there are some relevance between situations that give best values for the two criteria.

Example 3.6. In Table 2 we give the values of the above measures for the table given in Table 1. All the criteria select the variables 1 or 3 as a first test variable and can give the minimum tree indicated in Fig. 2.

4. Properties of the criteria

4.1. Judgement of constant tables

Since we compute values of the criterion for each variable i, it is efficient if we can decide whether we need further dividing of the table or not solely from the values of the criterion for all variables of the table. The most basic such recognition is whether the table is a constant or not. We present the conditions of constant tables for the three criteria.

Lemma 4.1. *$A_i = 0$ for all i if and only if f is a constant.*

Proof. Assume $fu^x i = a$. Since $A_i = 0$ for all i, complementing a single value $x(j)$ keeps the action invariant, because otherwise we have $A_j > 0$. Thus $fu^x i = a$ for any x. Hence f is a constant. □

Lemma 4.2. *If i is a unique essential variable of f then $H_l > 0$ for any $l \neq i$.*

Proof. Since $fl^0 = fl^1 \neq const$, $p_j^{l^0} \neq 0$ and $p_j^{l^0} \neq 1$ for some j. Hence $H_l > 0$ from (3). □

Lemma 4.3. *If f is a constant then $\sigma_{Bi}^2 = 0$ for all i. The converse is not true.*

Proof. Assume $f = a_j$. Then $\mu_j^i = \mu_T^i = 1/2$ for all i. Thus from Lemma 3.5 we have the assertion. It is easy to see that there is a function $f \neq cost.$ satisfying $\sigma_{Bi}^2 = 0$ for all i (cf. Example 4.1 below). □

Theorem 4.4. *The necessary and sufficient conditions for the constant judgement of the criteria Γ_A, Γ_H are $A_i = 0$, $H_i = 0$ for each i, while that for Γ_D is that there exists a unique variable j such that $w_j = 1$.*

Proof. The proofs for Γ_A and Γ_H are due to Lemma 4.1 and Lemma 3.2 and 4.2, respectively. The condition for Γ_D is trivial. □

Note 4.5. On the basis of Theorem 4.4 a practical procedure for constant recognition can be given by using A_i or H_i. If $max_i\ A_i = 0$ then f is a constant. Suppose $min\ H_i = 0$. If there is a l such that $H_l > 0$, then i is a unique essential variable of f. Otherwise f is a constant.

4.2. Rejection of nonessential variables (*nev*-free property)

We show that the criterion Γ_A do not select a nonessential variable, while Γ_H and $Gamma_D$ may do so. This is done by indicating that extremum values of the criteria Γ_H and $Gamma_D$ can not be given only by nonessential variables.

Lemma 4.6. *The discriminant criterion takes a minimum (worst) value $\sigma_{Bi}^2 = 0$ if i is a nonessential variable. The converse is not true.*

Proof. Assume that i is nonessential. Then we have the same number of $x_i = 0$ and $x_i = 1$ for each action a_j. Then obviously $N_j^{i^0} = N_j^{i^1} = N_j/2$. Hence from Lemma 3.5 we have the first assertion. □

Lemma 4.7. *The entropy H_i takes a maximum (worst) value if i is a nonessential variable. The converse is not true.*

Proof. Put $H_i := \sum_j h_j^i$, where $h_j := -(1/2)(p_j^{i^0} \log p_j^{i^0} + p_j^{i^1} \log p_j^{i^1})$. Since $p_j^{i^0} + p_j^{i^1} = 2(N_j^{i^0} + N_j^{i^1})/N = 2N_j/N = 2w_j$ does not depend on i. We can consider h_j^i as a function of $p_j^{i^0}$, $0 \leq p_j^{i^0} \leq 2w_j$. Then the function $h_j^i(p_j^{i^0})$ is non-negative and has value 0 at both boundaries, i.e. $h_j^i(0) = h_j^i(2w_j) = 0$. Further, it is upward convex and symmetric with respect to $p_j^{i^0} = p_j^{i^1} = w_j$. Hence it has a unique maximum when $p_j^{i^0} = p_j^{i^1}$. Thus H_i takes a maximum value if and only if $p_j^{i^0} = p_j^{i^1}$ for all j. □

Now we show that both converses of Lemma 4.6 and 4.7 are not true by examples. The reason for this is that $N_j^{i^0} = N_j^{i^1} = N_j/2$ for all j does not imply i to be nonessential. Indeed, in the examples below all the variables (either essential or nonessential) give respective tie values with respect to the both criteria H_i and σ_{Bi}^2. Thus, we are not sure that we do not select a worst variable (nonessential variable is a worst variable) so far as we are selecting a variable solely on the basis of the values of the criterion. The criterion Γ_A does not have this disadvantage. The unwelcome effect of this *nev* selection on the cost of the resulting trees is also discussed in [15].

We give 4-ary functions f in the following two examples (only 3-ary functions $f1^0$ are indicated since we set variable 1 (only) nonessential).

Example 4.8.

$$f1^1 := f1^0,\ f1^0(101) = f1^0(010) := b,$$
$$f1^0(x_2x_3x_4) = a \text{ for other } x_2x_3x_4.$$

We have $A_1 = 0$, $A_i = 2$ for $i = 2, 3, 4$. Also we have $N_a^{i^s} = 6$ and $N_b^{i^s} = 2$ for $i = 1, 2, 3, 4$ and $s = 0, 1$. Thus $\mu_j^i = 1/2$ for all i and j, leading to $\sigma_{Bi}^2 = 0$ for all i. Further, $p_a^{i^s} = 3/4$ and $p_b^{i^s} = 1/4$ for $s = 0, 1$ and $i = 1, 2, 3, 4$. Thus $H_i = -(3/4 \log 3/4 + 1/4 \log 1/4)$ $= 0.811$ for all i.

Example 4.9. The following $f1^0$ is a "parity" function.

$$f1^1 := f1^0,$$
$$f1^0(x_2x_3x_4) = \begin{cases} a & \text{if } x_2 + x_3 + x_4 = 0 \pmod 2, \\ b & \text{otherwise.} \end{cases}$$

We have $A_1 = 0$, $A_i = 4$ for $i = 2, 3, 4$. Also we have $\sigma_{Bi}^2 = 0$ and $H_i = 1$ for all i.

Thus in these cases Γ_D and Γ_H may select *nev* 1 while Γ_A does not. Note that this crucial *nev*-selection occurs *only when* $\sigma_{Bi}^2 = 0$ (a tie value) and $H_i = H_{\max} := (1/N)\sum_j N_j \log N_j + \log N$ (a tie value) for all i (cf. Note 3.3 and Lemma 4.7), since as far as

there exists i such that $\sigma^2_{Bi} \neq 0$ or $H_i \neq H_{\max}$no *nev*-selection occurs.

Hence we have:

Theorem 4.10. *Only the criterion Γ_A is nev-free, while the other criteria Γ_H and Γ_D may select a nonessential variable.*

4.3. Selection of a totally essential variable (*tev*-bound property)

Any *tev* is an optimal variable. So it is desirable for a criterion to select a *tev* whenever it exists. We show that this is true for Γ_A but not true for Γ_D and Γ_H. As we will see below, not only all *tevs* do not give extremum values but non-*tevs* also can give extremum values with respect to the both criteria Γ_D and Γ_H. Thus in a sense, they are not sensitive to total essentiality of a variable.

Theorem 4.11. *The criterion Γ_A is tev-bound, while Γ_H and Γ_D are not.*

Proof. This is derived from Lemma 3.1 and the example given below. □

In the following example all the variables (either totally essential or not) give tie values with respect to the both criteria Γ_D and Γ_H. Thus, we are not sure that we do select an optimal variable as far as we obey the criteria Γ_D or Γ_H. Indeed, they may select a non-*tev* 1 which is not optimal. On the contrary the criterion Γ_A does not have this disadvantage.

Example 4.12.

234	$f1^0$	$f1^1$
000	a	b
001	b	a
010	c	c
011	a	b
100	b	a
101	c	c
110	a	b
111	b	a

We have the following values of the criteria for each variable.

variable	essential.	optimal.	A_i	H_i	σ^2_{Bi}
1	non-tev	non-opt.	6	1.56	0.
2	tev	optimal	8	1.56	0.
3	tev	optimal	8	1.56	0.
4	tev	optimal	8	1.56	0.

The next example indicates that Γ_H and Γ_D do not give even unique values to totally essential variables.

Example 4.13.

123	$f4^0$	$f4^1$
000	a	c
001	b	c
010	b	a
011	a	b
100	b	a
101	a	b
110	a	c
111	b	c

We have the following values of the criteria for each variable.

variable	essential.	opt	A_i	H_i	σ^2_{Bi}
1	tev	optimal	8	1.56	0.
2	tev	optimal	8	1.56	0.
3	non-tev	non-opt.	6	1.5	1/48
4	tev	optimal	8	1.25	3/32

In fact we have $H_2 > H_3 > H_4$ and $\sigma^2_{B_2} < \sigma^2_{B_3} < \sigma^2_{B_4}$ and the variable 3 is not *tev* while the variables 2 and 4 are *tevs*. The criteria Γ_D and Γ_H select variable 4 (one of the optimal variables), while Γ_A may select any of *tevs* 1, 2 and 4.

4.4. Selection of a quasi-decisive variable (qdv-bound property)

Lastly we will show that the criteria Γ_D and Γ_H are *qdv*-bound while Γ_A is not, i.e. Γ_A may select a non-*qdv* even if there are *qdvs* among the variables of f. To show the *qdv*-bound property we must show that *qdvs* give an extremum (maximum or minimum) value of the criterion and, conversely, if there are *qdvs*, only they (any of them) give an extremum value.

First we need a notation for treating a table having *qdvs* which is used throughout this section. Assume that i is a 0-side-*qdv* ($fi^0 = a_1$), then f can be represented as follows:

i	k	actions
0 0 ... 0		
⋮	⋮	a_1
0 1 ... 1		
1 0 ... 0		
⋮	⋮	$\{a_{i_1}, \ldots, a_{i_K}\}$
1 1 ... 1		$K \geq 2$

So the condition that the variable i is a 0-side-qdv is:

$$\begin{aligned} N_1^{i^0} &= M, \\ N_j^{i^0} &= 0 \text{ for } j = 2, \ldots, K, \\ N_1^{i^1} &= N_1 - M, \\ N_j^{i^1} &= N_j \text{ for } j = 2, \ldots, K, \end{aligned}$$

where $M := N/2 = 2^{L-1}$ $(K \geq 2)$.

Let k be any variable of f $(k \neq i)$. Put

$$\begin{aligned} & m_j := N_j^{k^0}, \ n_j := N_j^{k^1}, \\ & (m_j + n_j = N_j, \sum_j m_j = \sum_j n_j = M), \quad (9) \\ & j = 1, \ldots, K. \end{aligned}$$

Now we have a lemma.

Lemma 4.14. *The variable k is also a qdv in f if and only if we have either $m_j = 0$ $(2 \leq j \leq K)$ or $n_j = 0$ $(2 \leq j \leq K)$.*

Proof. The necessity is easily seen from the condition that k is also a qdv. Since then we can represent the same table in the following form assuming that k is also decisive in 0-side:

	i				k	actions
fi^0	0	0	...	0	0	a_1 $(N/4)$
	⋮		⋮		⋮	
	0	1	...	1	0	
	0	0	...	0	1	a_1 $(N/4)$
	⋮		⋮		⋮	
	0	1	...	1	1	
fi^1	1	0	...	0	0	a_1 $(N/4)$
	⋮		⋮		⋮	
	1	1	...	1	0	
	1	0	...	0	1	$\{a_1, \ldots, a_K\}$ $(N/4)$
	⋮		⋮		⋮	
	1	1	...	1	1	

Then obviously $m_j = 0$ for $j = 2, \ldots, k$. The latter alternative occurs when k is 1-side-quasi-decisive. This is obtained by exchanging vectors of $x_k = 0$ and $x_k = 1$ in fi^1 so that the action a_1 covers $x_k = 1$. Conversely, assume that $m_j = 0$ for $j = 2, \ldots, K$. Then from (??) we have $n_j := N_j^{k^1} = N_j$ for $j = 2, \ldots, K$, $N_1^{k^0} = M$ and $N_1^{k^1} = N_1 - M$. This means that k is also a 0-side-qdv. □

We note that all the three criteria Γ_A, Γ_H and $Gamma_D$ give the same values, respectively, for all $qdvs$. This is easy to check for Γ_A and Γ_H, since activities are $A_i = \sum_{j=2}^{K} N_j = A_k$, and the entropy H_i is determined by the occurrence frequencies of actions in the two subtables fi^s (independently from its rule configurations). For discriminant criterion, this is not obvious. However, we have $N_1^{i^1} = N_1^{k^1} = N_1 - M$, $N_j^{i^1} = N_j^{k^1} = N_j$ $(2 \leq j \leq K)$ when both i and k are 0-side-decisive or $N_1^{k^1} = M$, $N_j^{k^1} = 0$ $(2 \leq j \leq K)$ when k is 1-side-decisive. In both cases, $\sigma_{Bi}^2 = \sigma_{B_k}^2 = (1/4)(N/N_1 - 1)$. Actually, we prove a more stronger result for each criterion in the sequel (i.e. every qdv gives an extremum value with respect to each criterion). We also prove that for activity criterion only a maximum value is not strictly attained by $qdvs$. This property makes the activity criterion not qdv-bound.

Lemma 4.15. *If i is a qdv, then A_i is a maximum. Converse is not true for $L > 3$, i.e. a maximum A_i is given not strictly by qdvs assuming there are qdvs, in other words, i may not be a qdv assuming that A_i is a maximum and there are qdvs in f.*

Proof. Assuming $fi^0 = a_1, fi^1 \neq const.$, we show $A_k \leq A_i$ for any k. Let D denote the set of rules having different actions from a_1 in f^{i^1} (the complement $\overline{D} = fi^1 \setminus D$ consists of rules having action a_1). Then the activity $A_i = |D| = \sum_{j=2}^{K} N_j$. Put $s := |D|$ for simplicity. For active k-pairs we are sufficient to consider only fi^1, since no active k-pair exists in fi^0. Assume that there are t active k-pairs within D. Then the other possibility of active k-pair is between D and $\overline{D}$ and the number of such k-pairs is at most $s-2t$. Hence total number of active k-pairs are at most $A_k \leq s-2t+t = s-t \leq A_i$. The equality holds if and only if $t = 0$, which is satisfied if (but not only if) the condition in Lemma 4.14 holds, i.e. k is also a qdv. For the second assertion we give an example below. □

In the following example we show that $A_k = A_i = s$ (a maximum among all variables) for a qdv i and for a non-qdv k. Thus Γ_A is not qdv-bound in this case.

Example 4.16. We represent a 4-ary function $f(x_1x_2x_3x_4)$ by giving $f1^0$ and $f1^1$:

$$\begin{aligned} & f1^0 := a, \quad f1^1(000) = b, \quad f1^1(011) = c, \\ & f1^1(110) = d, \quad f1^1(x_2x_3x_4) = a \text{ for other } x_2x_3x_4. \end{aligned}$$

We have $A_i = 3$ for all $i = 1, 2, 3, 4$ $(t = 0$ in all cases) but only the variable 1 is qdv.

We note that only for 2-ary tables $(L = 2)$ the converse of the lemma holds, i.e. the conditions $A_1 = A_2 =$ a maximum and variable 1 is qdv imply variable 2 also qdv.

Lemma 4.17. *If i is a qdv then H_i is a minimum. Conversely, if there are qdvs, then only qdvs give a minimum value of H_i.*

Proof. Let $fi^0 = a_1$ and k be any variable ($k \neq i$). We show $H_i \leq H_k$ and the equality holds (if and) only if k is also a *qdv*. Denote the number of rules having action a_1 and $x_k = 0$ in fi^1 by m_1' and similarly by n_1' for $x_k = 1$. That is

$m_1' :=$
$\quad |\{(x, f(x))$ having action a_1 and $x_i = 1,\ x_k = 0\}|,$
$n_1' :=$
$\quad |\{(x, f(x))$ having action a_1 and $x_i = 1,\ x_k = 1\}|.$

Since rules of fi^0 consist exactly the same number $(M/2)$ of vectors of $x_k = 0$ and $x_k = 1$, we have (cf. (9))

$$m_1 = M/2 + m_1', n_1 = M/2 + n_1' \ (N_1 - M = m_1' + n_1'). \tag{10}$$

From $p_j^{i^s} = N_j^{i^s}/M$ we have for i and k:

$$\begin{array}{ll} p_1^{i^0} = 1,\ p_j^{i^0} = 0 & \text{for } j = 2, \ldots, K, \\ p_1^{i^1} = (N_1 - M)/M,\ p_j^{i^1} = N_j/M & \text{for } j = 2, \ldots, K, \\ p_j^{k^0} = m_j/M & \text{for } j = 1, \ldots, K, \\ p_j^{k^1} = n_j/M & \text{for } j = 1, \ldots, K. \end{array}$$

Substituting these into (3), we have (hereafter all the summation is for $2 \leq j \leq K$ unless explicitly stated in other way):

$$\begin{aligned} s := -N(H_i - H_k) &= (N_1 - M)\log(N_1 - M)/M \\ &\quad + \sum (N_j \log N_j/M) - \sum_{j=1}^{K} (m_j \log m_j/M \\ &\quad + n_j \log n_j/M) \\ &= (N_1 - M)\log(N_1 - M) - (m_1 \log m_1 + n_1 \log n_1) \\ &\quad + M \log M + \sum (N_j \log N_j - m_j \log m_j - n_j \log n_j). \end{aligned}$$

Again, substituting $m_j = N_j - n_j$, $j = 1, \ldots, K$, we have

$$\begin{aligned} s &= (N_1 - M)\log(N_1 - M) \\ &\quad + M \log M - (N_1 - n_1)\log(N_1 - n_1) - n_1 \log n_1 \\ &\quad + \sum ((N_j - n_j)\log N_j/(N_j - n_j) + n_j \log N_j/n_j). \end{aligned}$$

Putting $t(x) = (N_1 - x)\log(N_1 - x) + x \log x$,

$$\begin{aligned} s &= t(M) - t(n_1) \\ &\quad + \sum ((N_j - n_j)\log N_j/(N_j - n_j) + n_j \log N_j/n_j). \end{aligned}$$

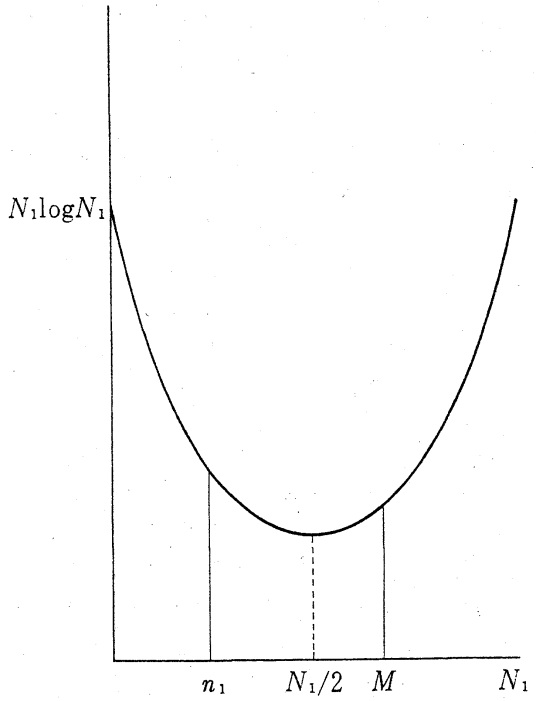

Figure 4: The function $t(x) = (N_1 - x)\log(N_1 - x) + x \log x$ $(0 \leq x \leq N_1)$.

From $N_j \geq n_j$ the last term is non-negative (≥ 0).

We show that k is a *qdv* if and only if $t(M) \leq t(n_1)$. The function $t(x)$, $0 \leq x \leq N_1$ is a monotone symmetric with respect to $x = N_1/2$ (cf. Fig. 4). Considering $n_1 \leq M$, $N_1/2 \leq M \leq N_1$, we have

$$n_1 \leq N_1/2 - (M - N_1/2) = N_1 - M \Leftrightarrow t(M) \leq t(n_1). \tag{11}$$

Substituting (10) into (11), we conclude $t(M) \leq t(n_1) \Leftrightarrow M/2 \leq m_1'$. Since there are another $M/2$ rules having action a_1 and $x_k = 0$ (among the rules of $x_i = 0$), this means that k is also 0-side-decisive, i.e. $fk^0 = a_1$ (a constant). Hence, if k is not a *qdv*, then $s > 0$. Thus, whenever a *qdv* exists, a *qdv* is selected (if there are many *qdvs* then any of them, because they give a tie value), since we select a variable i which gives a minimal value of H_i. □

Lemma 4.18. *If i is a qdv, then i gives a maximum value of σ_{Bi}^2. Conversely, if there are qdvs then only qdvs give a maximum value of σ_{Bi}^2.*

Proof. Instead σ_{Bi}^2 we can consider $B_i = \sum_j w_j(\mu_j^i)^2 =$

$(1/N)\sum_j (N_j^{i^1})^2/N_j$. Putting $N_j' := N_1 \cdots N_{j-1} \cdot N_{j+1} \cdots N_K$, we have $N \cdot N_1 \cdots N_K \cdot B_i = \sum_j N_j'(N_j^{i^1})^2$. Following the notation (10) for the variable k, consider

$$h := N \cdot N_1 \cdots N_K (B_i - B_k) = \sum_{j=1} N_j'((N_j^{i^1})^2 - n_j^2). \tag{13}$$

Substituting $N_1^{i^1} = N_1 - N/2$ and $n_1 = N/2 - \sum n_j$ into (13), we have the following:

$$\begin{aligned} h &= N_1'(N_1 - N/2)^2 + \sum N_j' N_j^2 \\ &\quad -N_1'(N/2 - \sum n_j)^2 - \sum N_j' n_j^2 \\ &= N_1 \cdots N_K (\sum_{j=1} N_j - N) + N_1'(N^2/4) \\ &\quad -(N_1'(N^2/4 - N\sum n_j + (\sum n_j)^2) - \sum N_j' n_j^2 \\ &= N_1' N \sum n_j - N_1'(\sum n_j)^2 - \sum N_j' n_j^2. \end{aligned}$$

Again, substituting $N = \sum_{j=1}^K N_j = N_1 + \sum N_j$ into this, we have

$$\begin{aligned} h &= N_1' N_1 \sum n_j \\ &\quad +N_1'(\sum n_j)\sum N_j - N_1'(\sum n_j)^2 - \sum N_j' n_j^2. \end{aligned}$$

Using $N_1' N_1 \sum n_j = \sum N_j' N_j n_j$, we finally have

$$\begin{aligned} h &= \sum N_j' N_j n_j \\ &\quad - \sum N_j' n_j^2 + N_1'(\sum n_j)(\sum (N_j - n_j)) \\ &= \sum N_j' n_j (N_j - n_j) + N_1'(\sum n_j)(\sum (N_j - n_j)) \\ &\geq 0. \end{aligned}$$

The equality holds strictly either $n_j = 0$ for $j = 2, \ldots, K$ or $N_j = n_j$ for $j = 2, \ldots, K$. This means that the equality $\sigma_{Bi}^2 = \sigma_{Bk}^2$ holds when and only when the variable k is also a *qdv* from Lemma 4.14. □

From Lemmas 4.15,4.17,4.18, we have the following theorem.

Theorem 4.19. *The two criteria Γ_H and Γ_D are qdv-bound, while Γ_A is not.*

5. Discussions and Conclusions

An efficient VSM (variable selection method according to a criterion) constructs a near optimum tree in the average much less computation than the worst case evaluation $O(L^2 2^L)$ with $O(L2^L)$ storage, where L is the number of variables of the table.

In this paper we have developed the three criteria for such VSM from three different standpoints: Γ_A (activity criterion) from combinatorial, Γ_H from entropy and Γ_D from discriminant analyses, for constructing an optimum tree in the sense of the number of nodes of the tree.

The three criteria have been examined with respect to the conditions which an optimum criterion should satisfy: rejection of nonessential variables (*nev*-free), selection of a totally essential variable and a quasi-decisive variable (*tev*-bound and *qdv*-bound properties; both *tev* and *qdv* are optimal variables). We have shown that Γ_A is *nev*-free, *tev*-bound but not *qdv*-bound, while the two criteria Γ_H and Γ_D are neither *nev*-free nor *tev*-bound but *qdv*-bound. It is hard to claim that one criterion is better than others only from these considerations. Consequently, a series of experiments was done, comparing the costs of these near optimum trees also with those of optimum trees. It shows that activity criterion is slightly better than others (1.03 versus 1.05 in terms of optimality coefficient) with comparable computation; the other two indicates practically identical performance.

Entropy of a variable is defined simply through occurrence frequencies of all actions regardless its structure in the conditional part and discriminant criterion inspects the rules in which the bit "1" is standing in the corresponding variable, while activity of a variable takes into account also the values of the other variables in the rule to some extent. This difference may have produced the slightly better coefficient for the activity.

An extension of the criteria to the case that testing a variable i incurs a certain cost C_i depending on the variable (general "description cost") may be to use a quantity given by dividing the value of the described criterion by C_i as a new criterion. Then, however, for such criteria one can not prove any of the formal properties which we have shown in this paper, although the properties that *nev* is a worst and *tev* and *qdv* are optimal variables remain valid.

VSM criteria for the case when the cost of a tree is defined as the average cost of testing can be obtained in the same way. We have observed a similar experimental result concerning the comparative performance of the three criteria. The combinatorial criterion for this cost has also been studied in more detail in [Miy89].

The newly introduced discriminant criterion Γ_D has an evident computational advantage over entropic criterion Γ_H, although the required numbers of basic operations (bit-test) remain the same order. This direction would provide us further development of the theory.

Acknowledgement

The first author wishes to dedicate this paper to the memory of the late Hitoshi Miyakawa.

References

[Bay73] Bayes A.J.: A dynamic programming algorithm to optimise decision table code. Australian Computer J. 5, 2 (May 1973), 77-79.

[Bud85] Budach L.: A lower bound for the number of nodes in a decision trees, Electron Inf. verarb. Kybern. EIK **21** (1985) 4/5, 221-228.

[Cha87] Chan, A. H.:Using decision trees to derive the complement of a binary function with multiple-valued inputs, IEEE Trans. on Comput., C-36, 2, 1987, 212-214.

[Fis36] Fisher R.A. The use of multiple measurements in taxonomic problems, Ann. Eugenics, 7, Part II, 179-188 (1936).

[Gar72] Garey, M.R.: Simple Binary identification problems. IEEE Trans. Comput. TC-21, 6, 588-590, June 1972.

[Gan73] Ganapathy.S., Rajaraman, V.: Information theory applied to the conversion of decision tables to computer programs. Comm. ACM, 16, 9, 532-539, Sept., 1973.

[HVMG82] Hartman , C.R.P., Varshney, P.K., Mehrotra,. K.G., Gerberich, C.L.: Application of information theory to the construction of efficient decision trees. IEEE Trans. on Information theory IT-28,4, 565-577, July 1982.

[HyR76] Hyafil, L., Rivest, R.L.: Constructing optimal binary decision trees is NP-complete. Comm. ACM 5, 1, 15-17, May 1976.

[Knu73] Knuth, D.E.: The art of computer programming, vol. 1, 2nd ed. Reading, Mass.: Addison-Wesley, 1973.

[Knu73] Knuth, D.E.: The art of computer programming, vol. 3, Reading, Mass.: Addison-Wesley, 1973.

[Lov85] Loveland, D. W,: Bounds for binary testing with arbitrary weights, Acta Informatica 22, 101-114, 1985.

[MiS80] Miyakawa, M., Sabelfeld, V.K.: On minimizations of size of logical schemes (in Russian). Theoretical basis of compiling (A. P. Ershov, ed.), 49-58, Novosibirsk State University, Novosibirsk 1980.

[MiO82] Miyakawa, M., Otsu, N.: Algorithms for constructing near-minimum total nodes decision trees from expanded decision tables (Japanese). TGEC IECE Japan, EC82-33, July 1982.

[Miy85] Miyakawa, M.: Optimum decision trees - an optimal variable theorem and its related applications -. Acta Informatica 22, 475-498, 1985.

[Miy89] Miyakawa, M.: Criteria for selecting a variable in the construction of efficient decision trees, IEEE Trans. Comput., to appear.

[MTG80] Moret, B.M.E., Thomason, M.G., Gonzalez, R.C.: The activity of a variable and its relation to decision table. ACM Trans. Prog. Lang. Syst. 2,4, 580-595, Oct. 1980.

[MTG81] Moret, B.M.E., Thomason, M.G., Gonzalez, R.C.: Optimization criteria for decision trees, Dept. of Computter Science, Collage of Engineering, University of New Mexico, Technical Report CS81-6, 1981.

[Mor82] Moret, B.M.E.: Decision trees and diagrams. Computing Surveys 14, 4, 593-623, Dec. 1982.

[Res66] Reinwald, L.T., Soland, R.M.: Conversion of limited-entry decision tables to optimal computer programs I: minimum average processing time. JACM 13, 3, 339-358, July 1966.

[ReS67] Reinwald, L.T., Soland, R.M.: Conversion of limited-entry decision tables to optimal computer programs II: minimum storage requirement. JACM 14, 4, 742-755, Oct. 1967.

[Sas85] Sasao, T: An algorithm to derive the complement of a binary function with multiple-valued inputs, IEEE Trans. on Comput., C-34, 2, February 1985, 131-140.

[ScS76] Schumacher, H., Sevcik, K.:The synthetic approach to decision table conversion. Comm. ACM 19, 6, 343-351, June 1976.

[Spr66] Sprague V. G.: On storage space of decision trees. Comm. ACM 9, 5, 319-319, May 1966.

[Ver72] Verhelst M.: The conversion of limited-entry decision tables to optimal and near optimal flowcharts: two new algorithms. Comm. ACM 15, 11, 974-980, November 1972.

[Weg84] Wegner I.: Optimal decision trees and one-time-only branching programs for symmetric Boolean functions. Information and Control **62**, 129-143 (1984).